`Join` Statements

Jake Eisaguirre

Table of contents

Introduction **2**

Left Join **2**
- **Example** . . . 2
- **Venn Diagram** . . . 4

Inner Join **5**
- **Example** . . . 5
- **Venn Diagram** . . . 6

Full Join **7**
- **Example** . . . 7
- **Venn Diagram** . . . 8

Introduction

When working with datasets, it is often necessary to combine them to gain a better understanding of the data. Joining is a common technique that merges two or more tables based on a common variable. In this tutorial, we will focus on three types of joins: left join, inner join, and full join.

Left Join

A left join (or left outer join) returns all the rows from the left table and the matching rows from the right table. If there is no match in the right table, the resulting table will contain null values. The left table is the first table in the join statement.

Example

Suppose we have two datasets, one with information about cities and another with information about the pollution levels in those cities. We can use a left join to merge the two datasets based on the city name.

```r
# Load libraries
library(dplyr)
library(kableExtra)
library(tidyverse)
```

```r
# Create data frame
city <- data.frame(
  city_name = c("New York", "Los Angeles", "Chicago", "Houston", "Phoenix"),
  state = c("NY", "CA", "IL", "TX", "AZ")
)
```

city_name	state
New York	NY
Los Angeles	CA
Chicago	IL
Houston	TX
Phoenix	AZ

```r
# Create data frame
pollution <- data.frame(
  city_name = c("New York", "Los Angeles", "Houston"),
  pollution_level = c(8, 6, 9)
)
```

city_name	pollution_level
New York	8
Los Angeles	6
Houston	9

```r
# Perform left join
lj_ex <- left_join(city, pollution, by = "city_name")
```

city_name	state	pollution_level
New York	NY	8
Los Angeles	CA	6
Chicago	IL	NA
Houston	TX	9
Phoenix	AZ	NA

The resulting table will contain all the rows from the city dataset and only the matching rows from the pollution dataset. For example, the row for Chicago will have a null value for pollution level since there is no match in the pollution dataset.

Venn Diagram

The left join can be visualized using a Venn diagram as follows:

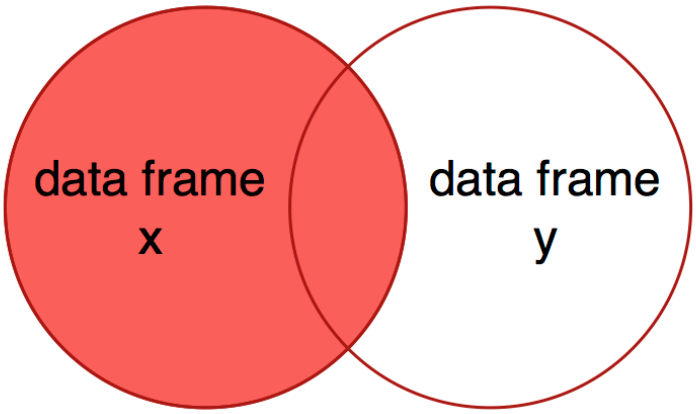

Figure 1: left_join

The resulting Venn diagram shows that all the rows from the City dataset (x) are included in the resulting table, but only the matching rows from the Pollution dataset (y) are included.

Inner Join

An inner join (or equi-join) returns only the rows that have matching values in both tables. The resulting table contains only the common values in both tables.

Example

Using the same datasets as in the left join example, we can perform an inner join to merge the two datasets based on the city name.

```
# Create data frame
city <- data.frame(
  city_name = c("New York", "Los Angeles", "Chicago", "Houston", "Phoenix"),
  state = c("NY", "CA", "IL", "TX", "AZ")
)
```

city_name	state
New York	NY
Los Angeles	CA
Chicago	IL
Houston	TX
Phoenix	AZ

```
# Create data frame
pollution <- data.frame(
  city_name = c("New York", "Los Angeles", "Houston"),
  pollution_level = c(8, 6, 9)
)
```

city_name	pollution_level
New York	8
Los Angeles	6
Houston	9

```r
# Perform inner join
ij_ex <- inner_join(city, pollution, by = "city_name")
```

city_name	state	pollution_level
New York	NY	8
Los Angeles	CA	6
Houston	TX	9

Venn Diagram

The inner join can be visualized using a Venn diagram as follows:

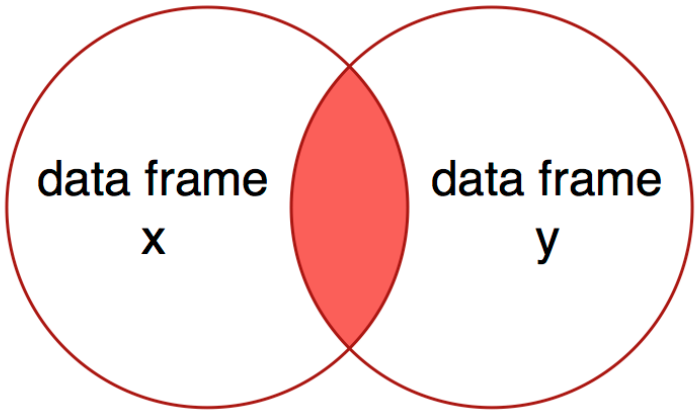

Figure 2: inner_join

The resulting Venn diagram shows that only the common values in both datasets are included in the resulting table.

Full Join

A full join (or full outer join) returns all the rows from both tables, including the rows that have no matching values in the other table. If there is no match in one of the tables, the resulting table will contain null values.

Example

Using the same datasets as in the left join and inner join examples, we can perform a full join to merge the two datasets based on the city name.

```r
# Create data frame
city <- data.frame(
  city_name = c("New York", "Los Angeles", "Chicago", "Houston", "Phoenix"),
  state = c("NY", "CA", "IL", "TX", "AZ")
)
```

city_name	state
New York	NY
Los Angeles	CA
Chicago	IL
Houston	TX
Phoenix	AZ

```r
# Create data frame
pollution <- data.frame(
  city_name = c("New York", "Los Angeles", "Houston", "Tampa"),
  pollution_level = c(8, 6, 9, 4)
)
```

city_name	pollution_level
New York	8
Los Angeles	6
Houston	9
Tampa	4

```r
# Perform full join
fj_ex <- full_join(city, pollution, by = c("city_name"))
```

city_name	state	pollution_level
New York	NY	8
Los Angeles	CA	6
Chicago	IL	NA
Houston	TX	9
Phoenix	AZ	NA
Tampa	NA	4

The resulting table will contain all the rows from both datasets. If there is no match in either dataset, the corresponding columns will have null values.

Venn Diagram

The full join can be visualized using a Venn diagram as follows:

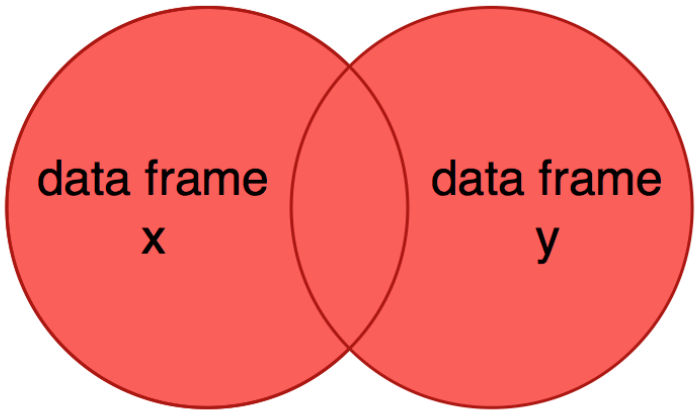

Figure 3: full_join

The resulting Venn diagram shows that all the rows from both datasets are included in the resulting table, with null values for any non-matching rows.